IMPACT ANALYSIS OF FIRES ON PUBLIC HEALTH AND SAFETY

DATA 512 FINAL REPORT
PRERIT CHAUDHARY

Introduction

The number of wildfires has been on the rise over the last few years and can have a huge impact on several aspects of a city. The number of fires have increased over the last few years and have a significant impact on the environment. Urban and rural landscapes are impacted in many ways. Urban landscapes suffer a more significant impact due to the density of people living in the city which affects both the quality of human lives and overall infrastructure. Beyond the impact of fires on an immediate level, they also have long term consequences on healthcare, environment, and economic aspects of a city. People living in the city may see an increase in respiratory issues, cardiovascular issues, and road accidents due to low visibility. Apart from their effect on health, fires also have repercussions on the economic conditions through property damage, disruption of businesses, agricultural losses, and tourism. The city council needs to have a proper plan to address all these concerns and make people aware of the measures that they have taken to overcome these losses. There is also a need to see if a certain group of people are more vulnerable to this and introduce new measures to protect these individuals.

Derby is one of the largest suburbs in Wichita, Kansas. The population of the city is 24,000 with the city being known for its green spaces and clean air. However, there has been an increasing number of cases of wildfires over the last few years. The aim of the analysis is to understand the impact of these fires on the health and safety of its residents. The analysis involved understanding the correlations between fires and mortality rates due to respiratory diseases, cardiovascular health, and road accidents over the last few years. Smoke fires can introduce different particulate matter and gases into the air which can cause several health issues and visibility reduction which can have short term and long-term consequences. This can impact not just the residents across all demographics. The analysis also involves checking the existence of demographic disparity by evaluating if a certain age group or ethnicity has had a significant effect on its life expectancy or mortality rates over the last few years.

The approach adopted to conduct this analysis is human centered so that the insights and recommendations from this analysis lead to policy changes that improve the quality of life of the residents living in Derby. There is a need for deep diving into the data to investigate trends across different communities and groups other than the overall trends. This will allow the city council to focus on the individuals most impacted by the fires and lead to actionable changes. An example of changes can be to improve access to healthcare for the group of people most affected by fires and identify if guidelines can be changed for certain professions that could be causing health issues. The analysis is done in an interpretable and explainable manner to ensure transparency and reliability of the findings.

The end goal is to be able to inform policy makers, city managers, city councils, or other civic institutions, to make an informed plan for how they could or whether they should make plans to mitigate future impacts from wildfires.

Background/Related Work

There has previously been credible research published on understanding the impact of fires on health and safety. Few of the paper referred before conducting this analysis are :

- Respiratory Impacts of Wildland Fire Smoke: Future Challenges and Policy Opportunities. An Official American Thoracic Society Workshop Report [1]

  This paper talks about a workshop that was convened to examine the science of wildland fire incidence and fire management, effect of wildfire smoke on respiratory and cardiovascular health, strategies aimed at addressing these health risks, and subsequent actions taken.

- Differential respiratory health effects from the 2008 northern California wildfires: A spatiotemporal approach [2]

  This paper researched on the impact of a large wildfire complex in northern California that occurred in the summer of 2008 where daily wildfire-PM exposure was modeled using a data-adaptive machine learning approach with spatiotemporal data sets. The main finding was that for each 5µg/m3 increase in wildfire-PM2.5, the risk of emergency department visits for asthma [RR=1.06, 95% CI=(1.05, 1.07)] and COPD [RR=1.02 (95% CI=(1.01, 1.04)], and hospitalizations [RR=1.07, 95% CI=(1.05, 1.10)] increased.

- Responses to Wildfire and Prescribed Fire Smoke: A Survey of a Medically Vulnerable Adult Population in the Wildland-Urban Interface, Mariposa County, California[3]

  This paper studied the impact of smoke fire on the community health experience, adaptive capacity, and attitudes to reform public health guidance. They conducted a survey of vulnerable individuals in a rural, high fire risk county (N = 106, 76% > 65 years) regarding wildfire and prescribed smoke health effects, health protective actions, information needs; and support for fire management policies. The comments were reviewed for context which reported that more than half (58%) of participants reported health impacts from wildfire smoke and 26% experienced impacts from prescribed fire smoke.

- Modeling the impact of wildfire smoke on driving speed [4]

  This paper aimed at analyzing the effectiveness of evacuation as drivers would reduce their speed in relation to the optical density on the road due to smoke fire. This would explicitly represent the impact of reduced visibility conditions on traffic evacuation flow.

The project was divided into two halves where the first part of the project was focused on the smoke estimate calculation for Derby, Kansas and the second part was extending that analysis to see the impact of fires on Public Health and Safety. In the first part of the project, various techniques were explored such as Linear Regression, ARMA (Autoregressive Moving Average), ARIMA (Autoregressive Integrated Moving Average) models but ended up using SARIMA(Seasonal Autoregressive Integrated Moving Average (SARIMA)) for my forecasting model. This is because SARIMA explicitly incorporates seasonality into the model, which makes it a great fit for the time series data available to us. This helps in utilizing the recurring patterns at regular intervals to make more robust predictions. Also, SARIMA has a more flexible framework which can tune autoregressive and moving average components which allows the model to capture nonlinear patterns present in time series data.

After reading through these research papers, I extended my research on Healthcare and thus focused my analysis on the effects of smoke fires on people's health and overall safety. The analysis mainly focused on the following aspects :

- Respiratory Health Issues : Exposure to wildfire smoke can exacerbate respiratory conditions such as asthma, bronchitis, and chronic obstructive pulmonary disease (COPD)
- Cardiovascular Effects : Fine particles in smoke can contribute to cardiovascular problems, potentially increasing the risk of heart attacks and other cardiovascular issues.
- Visibility Reduction : Thick smoke can reduce visibility, posing hazards for drivers and increasing the risk of accidents on roads.

The analysis would revolve around answering the following questions around the impact of smoke fires on residents in Derby, Kansas -

Mortality Rates for Respiratory Diseases, Cardiovascular Diseases and Road Accidents

- Possible correlations between increase in smoke fires and mortality rates because of the diseases
- Which respiratory diseases are shown to have a significant increase in mortality rates?
- Observe trends in road accidents at the time of smoke fires due to low visibility.

Demographic Disparity

- Identifying age groups or ethnicities most vulnerable or impacted by health conditions.
- Observe change in life expectancy across different demographics.

Methodology

Data Collection and Configuration

The first step in the pipeline involved fetching data from the provided data sources. The first dataset was the USGS Wildland Fire Combined Dataset maintained by the US Geological Survey, available in GeoJSON format.

For the extension phase of the project, data was sourced from the GHDx website which is the world's most comprehensive catalog of surveys, censuses, vital statistics, and other health-related data. The data was available in csv format.

Detailed data profiling for the datasets is provided in the Data Source subsection.

Exploratory Data Analysis

The first half of the project focused on building the entire data pipeline to create data in such a format that can be used for any downstream analysis. This involved the following steps:

- Using Python to load, filter and aggregate data to create intermediate datasets which can be used for analysis. This included filtering fires within 1250 miles of Derby and aggregating this for the years between 1963 to 2023.
- The missing data was imputed, and data quality was ensured to ensure fires were captured during the season

Calculating Fire Estimate and Forecasting Model

A well-defined formula was used for calculating the smoke fire across years. This was then compared with the AQI aggregated over the years to do an initial analysis in the first phase of the project. The formula involved :

- Size of the fire in GIS acres

  The size of the fire directly affects the amount of smoke produced and released. The larger the size of the fire, greater would-be particulate matter released. The formula also accounts for metrics and converts the acres into square miles using the standardized conversion (0.0015625)

- Distance of the fire

  The distance of the fire also has a direct impact on its effect. Lesser the distance from the fire, more the smoke can penetrate and impact the health and safety.

The final formula for the smoke estimate was as follows :

Smoke Estimate = (GIS Acres * 15.625) / Distance

SARIMA(Seasonal Autoregressive Integrated Moving Average (SARIMA)) was used for the forecasting model. This is because SARIMA explicitly incorporates seasonality into the model, which makes it a great fit for the time series data available to us. This helps in utilizing the recurring patterns at regular intervals to make more robust predictions. Also, SARIMA has a more flexible framework which can tune autoregressive and moving average components which allows the model to capture nonlinear patterns present in time series data. This was then used for predicting smoke between 2024- 2049 along with the confidence intervals for the analysis. The AQI in Derby has remained in the green zone which is indicative of the high tree coverage in the area and hence the second phase of the project was focused on deciphering the impact of smoke fires.

Statistical Analysis

The second part of the project involved understanding the impact of smoke fires on the health of residents living in Derby, Kansas. The area of focus is Health care as it is important to identify key trends and patterns to bring about a change for the betterment of the residents living there. The analysis would involve combining the previously built predictive model for smoke fires in the near future with the new data sources identified to study the effects of smoke fires because doing so helps us understand how they affect people's health, the environment, and overall safety.

The analysis mainly focused on the following aspects :

- Respiratory Health Issues : Exposure to wildfire smoke can exacerbate respiratory conditions such as asthma, bronchitis, and chronic obstructive pulmonary disease (COPD)
- Cardiovascular Effects : Fine particles in smoke can contribute to cardiovascular problems, potentially increasing the risk of heart attacks and other cardiovascular issues.
- Visibility Reduction : Thick smoke can reduce visibility, posing hazards for drivers and increasing the risk of accidents on roads.

This was done using the following statistical tests to check for significance :

- **Pearson Correlation Coefficient and Spearman's Rank-Order Correlation**

  Statistical methods were used to quantify the strength and direction of relationships between two variables. Pearson Correlation has an assumption that the data is normally distributed and hence I also used Spearman's Rank-Order Correlation so that there is no assumption on data distribution.

- **Analysis of Variance (ANOVA)**

  ANOVA is an extension of the t-test for more than two groups. It assesses whether there are any statistically significant differences in means among the groups. This was used to checked if there is an actual difference in average mortality rates across communities.

Human Centered Approach

Reproducibility

- The methodology is completed in a reproducible manner and only interpretable analysis is selected for the results. A few of the pointers include :
  - Documented Code : The Code includes Markdown and inline comments for Part 1 and extension analysis.
  - Data Dictionary: All data sources have been provided with a few important columns that were used for analysis.
  - A hierarchical project repository has been maintained to make it easier to navigate and skim through the project highlights.

Biases in data

- The data contains all the diverse representations for age and race across the county to ensure there is no misrepresentation of trends. A few checks that were made to ensure no bias are as follows :
  - Completed a quality check to look for any discrepancies in terms of age, race, gender, or any other demographic variables. It was ensured every group was represented equally so that there was no misrepresentation of patterns.
  - It was validated that the source of the data had enough data which fairly represents the diversity of the population under study.
  - There was special attention given to social context while choosing the parameters(Comparative Analysis across gender, age, or race) for which analysis was done.

Fairness , accountability, transparency and ethics (FATE)

- The analysis is performed using explainable techniques to make the results intuitive and reliable.
  - Ethical Considerations - Special attention was given while shortlisting the results to be presented such as comparing road accidents by gender and which ensured that the analysis aligned with ethical standards.
  - It was ensured that the analysis results are explainable with precise explanations of statistical methods and model choices with their reasoning interpretation.
  - The data uses aggregated data to mask PII (Personally Identifiable Information), and it was ensured that the data contains no personal information

Findings

There were patterns that were visible in the initial part of the analysis which can be seen through the following three visualizations :

- A histogram showing the number of fires occurring every 50-mile distance from Derby up to 1250 miles.

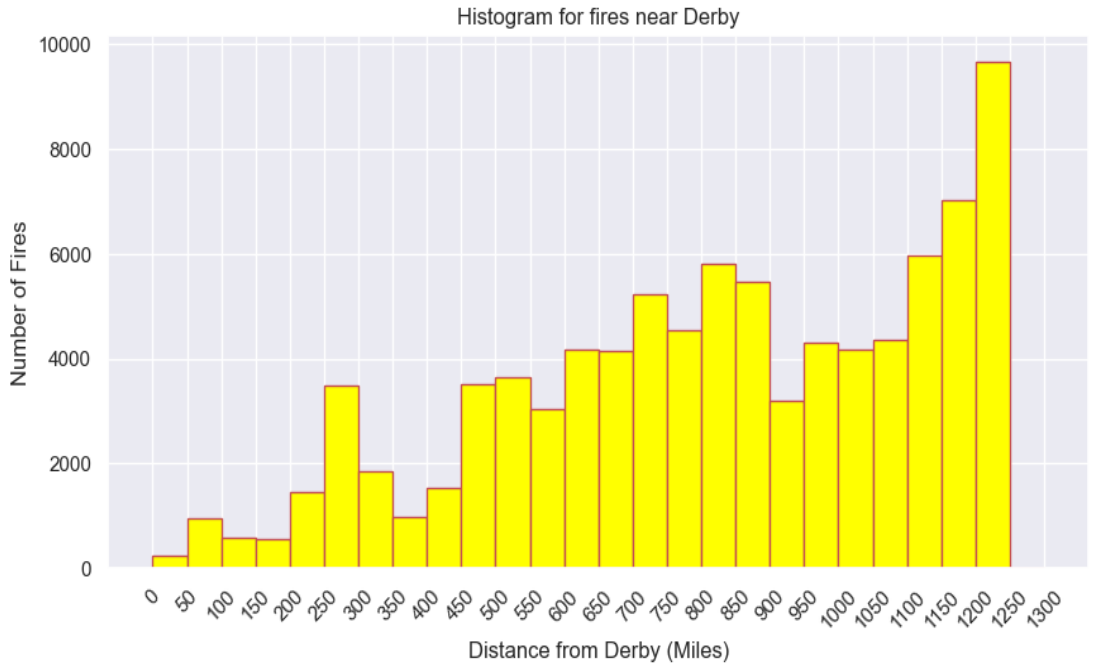

Fig 1: Fires within 1250 miles of Derby

A time series graph of total acres burned per year for the fires occurring within 1250 miles from Derby, Kansas

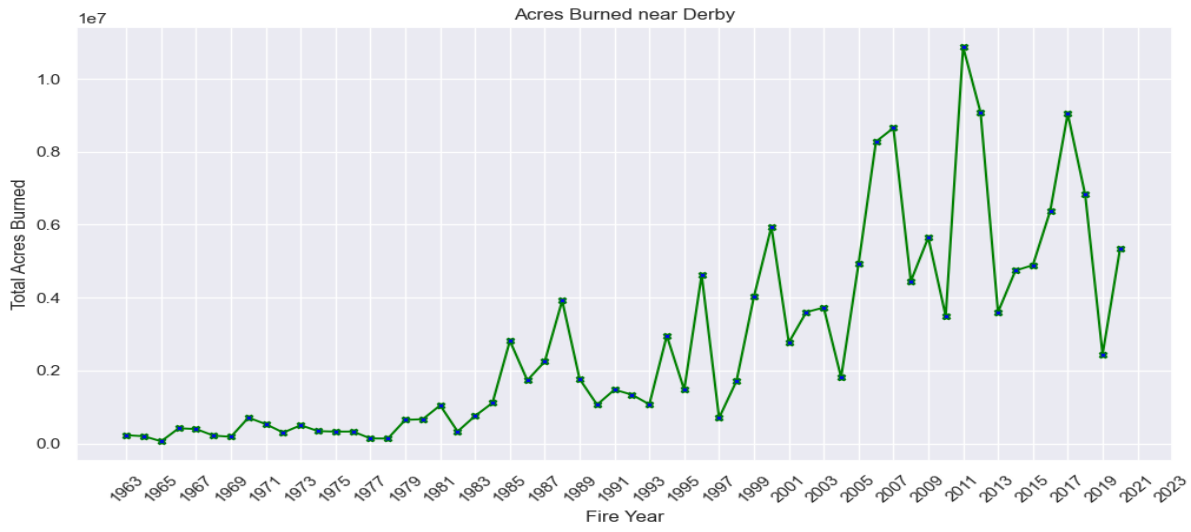

Fig 2: Acres burned over the years

- A time series graph containing and comparing the fire smoke estimate and the AQI estimate for Derby over the last few years

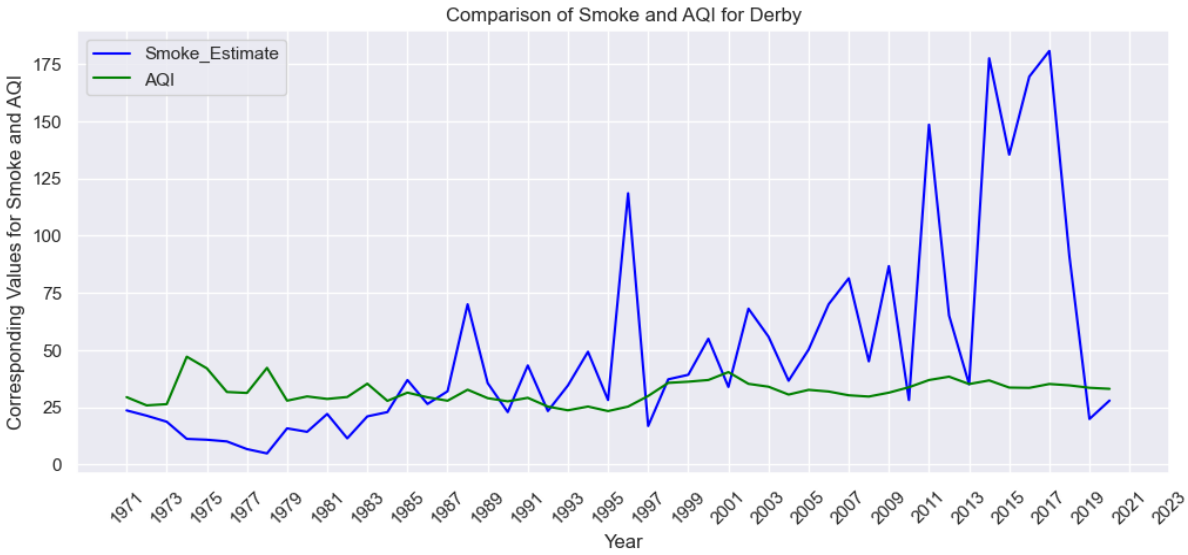

Fig 3: Comparison of AQI and Smoke Estimate over the years

SARIMA model was then used for forecasting for the years 2024-2049. The first step to implement that model was to do a normality test for which three normality tests were done which determined how likely a data sample is from a normally distributed population using p-values.

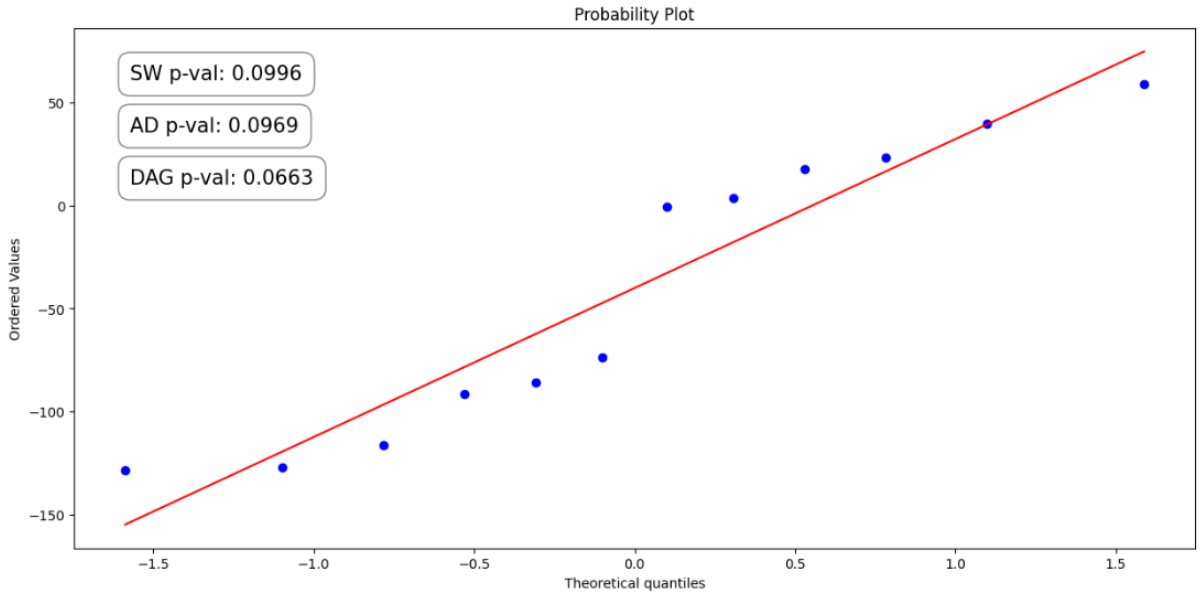

Fig 4: Tests for Normality

All three tests returned a p-value greater than the alpha value of 0.05. That means that the null hypothesis cannot be rejected and it is likely that the data points come from a normal distribution.

The model was then fit by doing a train and validation split on the data available to test how the model is doing on the already known result.

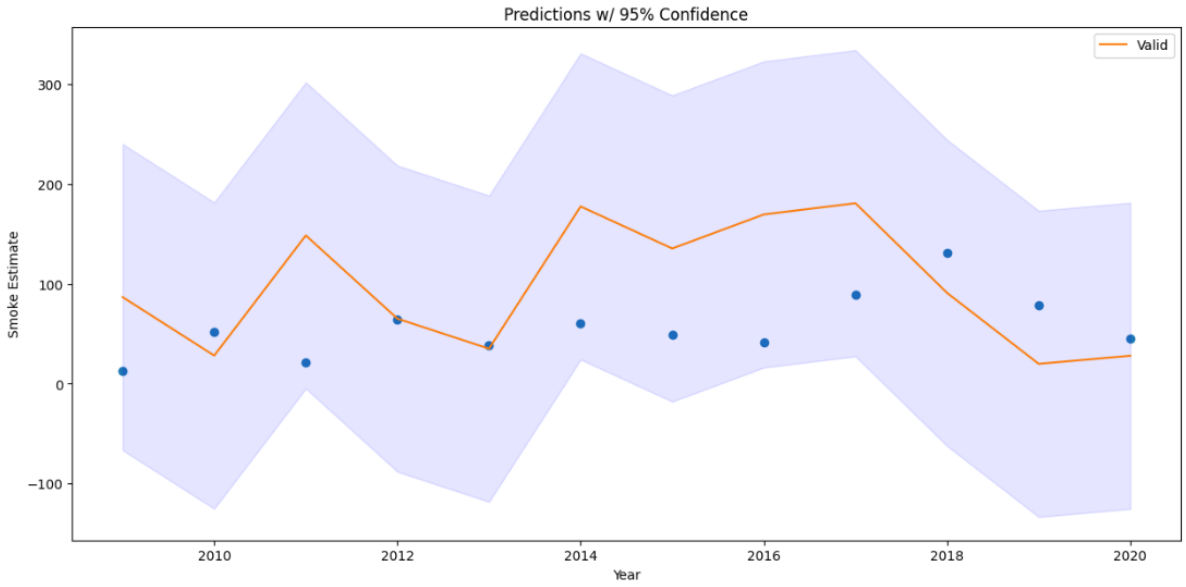

Fig 5: Predictions on Validation Set

The above figure shows the predictions for Validation Set in the below plot with confidence intervals where the orange line is displaying the true values, and the blue points are indicating the predicted values. Predictions were then made using this model for the Test Set for values between 2024- 2049 which can be seen in the below plot.

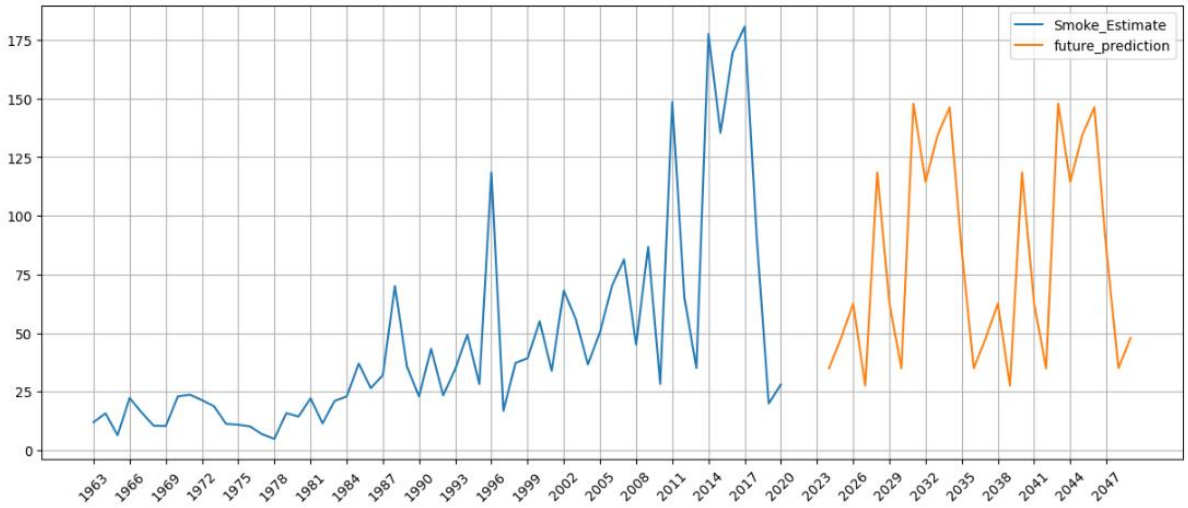

Fig 6: Prediction using SARIMA for 2024-2049

The key takeaways and findings from the first phase of the project included :

- One major insight that came out from the data was that there was no data available beyond 2020 for fires within 1250 miles from Derby, Kansas. One possible hypothesis for this can be the commencement of the pandemic which had the world in a lockdown. Derby, being one of the largest suburbs in Wichita is known for its green patches which is reflected by the smaller number of fire cases being reported in the proximity of Derby.
- However, the cases increase significantly as we move away which might be indicative of commercialization and industrialization. It can also be seen that over the last few years, there

has been a significant rise in the acres of land being burned all over the world due to fires due to several factors such as climate change, industrialization, commercialization etc.

- Derby faced a similar trend where the acres of land peaked in 2011. However, there is always a single event that brings about a change and ever since 2011, the area of land being burned has decreased which validates that proper measures are being taken to control the situation.

The extension plan which was the second phase of the project involved using the Smoke Estimate to see how it impacts Public's Health and Safety. There were 3 main findings from the analysis :

1. **Comparative Analysis of Mortality Rates for Respiratory Conditions between Males and Females**

   The below plot shows the increasing mortality rates across different respiratory conditions between males and females. The plot has a line plot representing the Smoke Estimate for Sedgwick County in the background in pink.

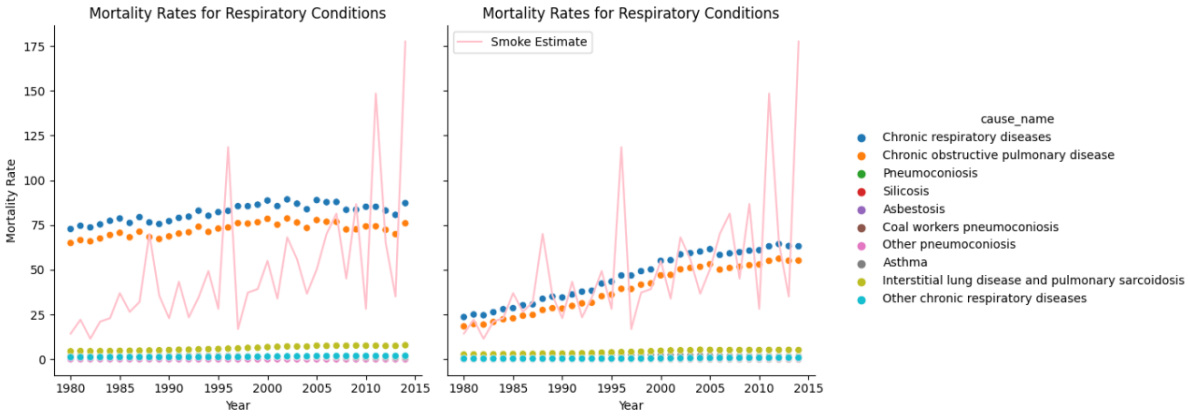

Fig 7: Mortality Rates for Respiratory Conditions between Males and Females

To further validate the trend seen in the visualization, I checked the percentage change in mortality rate for all respiratory diseases between males and females which can be seen in the below plot

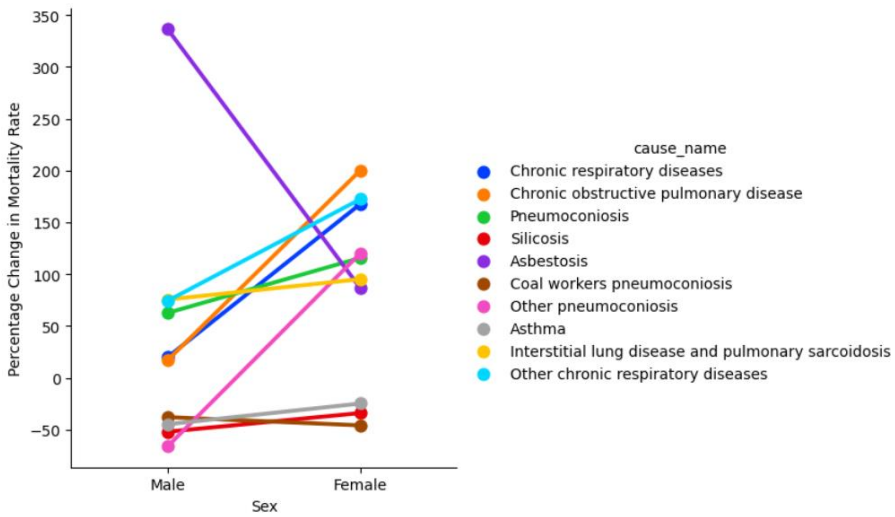

Fig 8: Percentage Change in mortality rates for respiratory conditions between males and Females

I then validated my Hypothesis seen in the above two visualizations by calculating the correlation between Increasing Mortality Rate in Females for Chronic obstructive pulmonary disease and Smoke Estimate over the last few years. I calculated both the Pearson correlation coefficient and Spearman correlation coefficient. Pearson correlation coefficient has an assumption that the data is normally distributed and hence Spearman correlation coefficient was also calculated which has no assumptions¶

```
Pearson Correlation Coefficient: 0.5477493747167843
P-Value: 0.0006595317716111251


Spearman Correlation Coefficient: 0.6630252100840336
P-Value: 1.4236497356140295e-05
```

Fig 9: Correlation Tests

It can be observed that p-value is significant for both tests even when there is no assumption on normality.

2. Comparative Analysis of Mortality Rates for different ethnicities for Cardiovascular Diseases

I noticed that mortality rates were higher for AIAN (American Indian and Alaska Native) for both respiratory and cardiovascular health conditions. I just plotted the trend for cardiovascular health to emphasize this which can be seen in the below plot

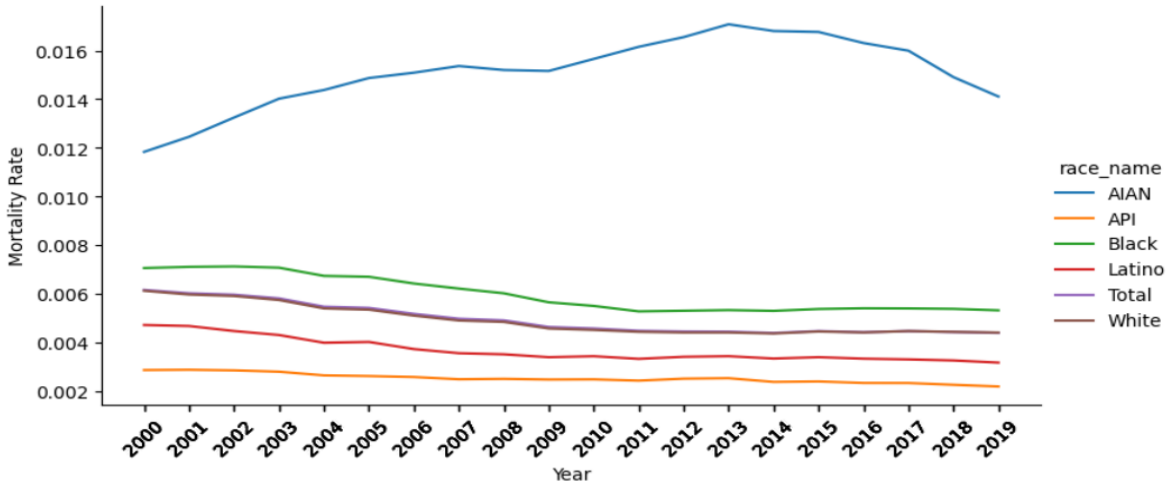

Fig 10: Mortality Rates for different ethnicities for Cardiovascular Diseases

Doubling down on this, I then wanted to further confirm this pattern by checking the Life Expectancy at birth for different ethnicities which can be seen in the below plot.

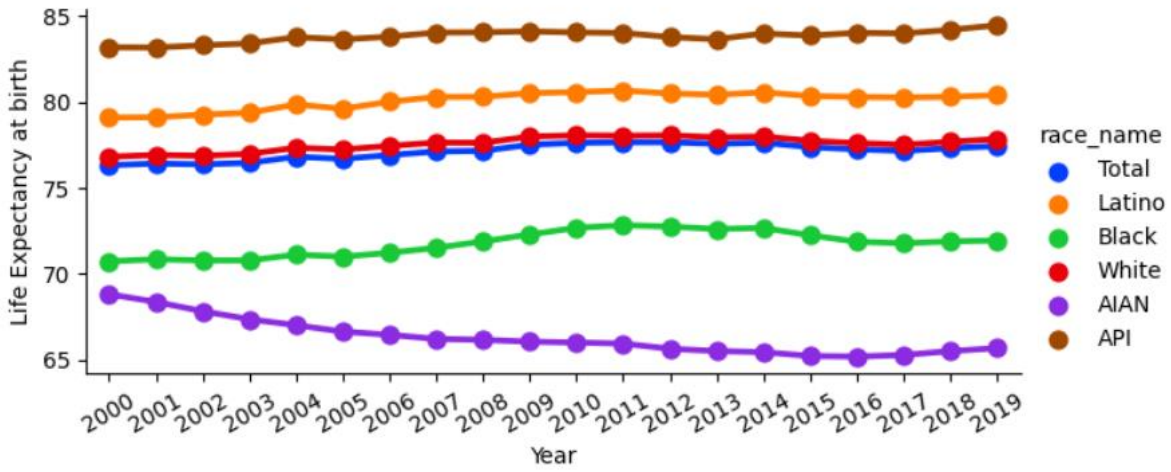

Fig 11: Life Expectancy for different ethnicities for Cardiovascular Diseases

My hypothesis was affirmed as the Life Expectancy for AIAN has been on decline over the last few years. The next step was to check if there was any statistical significance to the difference in average mortality rate for different races. I used ANOVA (Analysis of Variance) as there were more than two groups to check for differences between average mortality rates of different groups.

	sum_sq	df	F	PR(>F)
race_name	0.002045	5.0	1723.636744	3.035669e-101
year	0.000007	1.0	28.369780	5.512251e-07
year:race_name	0.000037	5.0	31.517899	1.136085e-19
Residual	0.000026	108.0	NaN	NaN

Fig 12: ANOVA results

The results of ANOVA showed that the interaction effect (race _name and year) is statistically significant ($p < 0.05$), indicating that there is evidence of an interaction effect and thus there is a significant difference in the average mortality rate across different ethnicities

3. **Comparative Analysis of Road Accidents and Transport Injuries with the Smoke Estimates over the years across different Age groups**

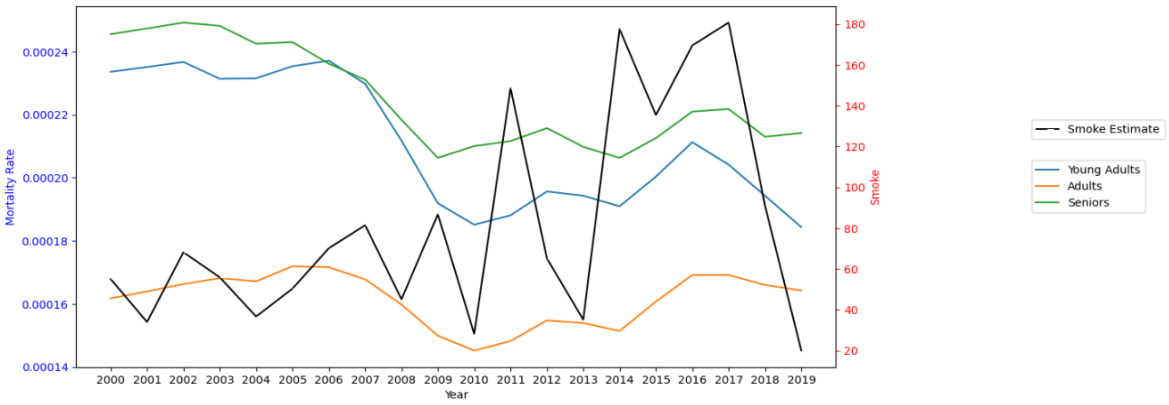

Fig 13: Road Accidents for different Age Groups over the years

I plotted the Average Mortality Rate for different Age groups along with the Smoke Estimate for those years in the background to see how increase in smoke has an impact on the mortality rate due to road accidents which can be seen in the above plot.

There are very local patterns where the mortality rate is high across all age groups whenever there is an increase in smoke that year. I delved deeper to examine trends among seniors, considering that reaction times may be affected by age-related factors, potentially leading to slower reflexes. To validate this, I then checked for statistical significance in terms of Correlation between increase in Mortality rates due to road accidents and Smoke Estimate especially for the Senior Age Category.

```
Pearson Correlation Coefficient: -0.3632132603672961
P-Value: 0.11547330752707058


Spearman Correlation Coefficient: -0.2270676691729323
P-Value: 0.3356744310134179
```

Fig 14: Correlation results for Road Accident and Smoke Estimate

Though the correlation came out to be Negative between Road Accidents for Seniors and Smoke Estimate. However, an important thing to notice is that the p-value is not less than 0.05 and hence this correlation is not statistically significant¶.

Discussion/Implications

The analysis done through this project shows the need for the city council to intervene and address the impact of increasing smoke fires in Derby, Kansas. The key findings and takeaways from the project are as follows :

- Found positive correlation between Smoke Fire Estimate and Increasing Mortality Rates in Females:

  Null hypothesis : No true correlation between smoke fire and mortality rate

  p-value is significant for both tests even when there is no assumption on normality

```
Pearson Correlation Coefficient: 0.5477493747167843
P-Value: 0.0006595317716111251


Spearman Correlation Coefficient: 0.6630252100840336
P-Value: 1.4236497356140295e-05
```

Fig 15: Correlation results for Respiratory conditions and Smoke Estimate

- ANOVA between mortality rate , race and year for cardiovascular disease

  The interaction effect (race_name and year) is statistically significant ($p < 0.05$), indicating that there is evidence of an interaction effect

	sum_sq	df	F	PR(>F)
race_name	0.002045	5.0	1723.636744	3.035669e-101
year	0.000007	1.0	28.369780	5.512251e-07
year:race_name	0.000037	5.0	31.517899	1.136085e-19
Residual	0.000026	108.0	NaN	NaN

Fig 16: ANOVA results

- Negative Correlation between Road Accidents for Seniors and Smoke Estimate

  However, there is an important thing to notice that the p-value is not less than 0.05 and hence this correlation is not statistically significant.

```
Pearson Correlation Coefficient: -0.3632132603672961
P-Value: 0.11547330752707058


Spearman Correlation Coefficient: -0.2270676691729323
P-Value: 0.3356744310134179
```

Fig 17: Correlation results for Road Accidents for Seniors and Smoke Estimate

These findings indicate that there is an urgent need for the city council to intervene and take actionable steps for the improvement of public health and safety in the city of Derby, Kansas. The results lay down a path for key areas that can be focused to streamline their approach to address the effect of fires on residents in Derby. This would allow them to strategically build a plan towards allocating the required number of resources and finances towards improving the quality of life.

Recommendation and Implications

The results provide insights into how smoke fires affect various aspects, including respiratory conditions, cardiovascular diseases, and reduced visibility leading to road accidents.

- The county should implement necessary measures to address the rising mortality rates related to respiratory conditions among females especially for Chronic obstructive pulmonary disease and Chronic respiratory diseases. This can be done through the following steps :

  - Launch targeted health education campaigns to raise awareness about the risks of respiratory conditions, emphasizing the importance of early detection, symptom recognition, and seeking medical attention.
  - Implement regular screening programs, especially for vulnerable populations, to detect respiratory conditions at an early stage. Early detection allows for timely intervention and better management of chronic diseases.
  - Improve access to healthcare services, ensuring that females have easy access to medical facilities, clinics, and specialists for respiratory health check-ups and treatment.

- Special attention is required for the AIAN ethnicity group where there has been an increase in mortality rates for respiratory and cardiovascular diseases. Life expectancy has also shown a visible decline for AIAN over the last few years.
    - Understand if the AIAN community is more involved in jobs that are more prone to health issues so that appropriate steps and strategies can be launched to address occupational health disparities.
    - Implement robust programs that can reform and enforce regulations in industries where the AIAN community is employed. Workers need to be educated on potential health hazards and provided with appropriate protective equipment.
    - Improve access to quality education by investing in schools and educational resources within or near AIAN communities. Enhance cultural competence in educational curricula to make them more inclusive.
    - Work with community leaders to act as liaisons between healthcare providers and community members to empower them with healthcare services, including preventive care, screenings, and chronic disease management which can foster trust and cultural understanding.

- It is advisable to adjust driving guidelines, particularly during periods of increased smoke fires, to ensure road safety.
    - Launch public awareness campaigns to educate people about the risks associated with reduced visibility due to smoke fires.
    - Utilize emergency alert systems to notify drivers of hazardous conditions caused by smoke fires.
    - Introduce variable speed limits that can be adjusted dynamically based on current visibility conditions.

Limitations

My analysis contains caveats and limitations which must be considered while going through the results and findings. The limitations extend across data exploration, model training or interpreting the predictions of the model.

- The following recommendations are based on the results of the correlation and other statistical tests conducted on the data. These results do not imply causation.
- The insights drawn from the analysis are only for the period data was available.
- The calculation for smoke estimates only contains the size and distance of the fire and might not be the best approximation.
- The forecasting model SARIMA forecasting model only performs well when the data has non-seasonal trends or structural changes and hence there might be better alternative models that can be explored.
- The analysis is limited by the variables included in the study. There might be other confounding variables which would directly impact the trends and a more exhaustive analysis is necessary.
- If the sample of demographic data is not representative of the broader population, the results may not be generalizable.
- Both correlation tests and ANOVA may overlook important contextual factors that influence demographic patterns. It's crucial to interpret results within the broader social, economic, or cultural context.

Conclusion

The analysis was focused on the following research questions which were answered through the available data:

Mortality Rates for Respiratory Diseases, Cardiovascular Diseases and Road Accidents

- Possible correlations between increase in smoke fires and mortality rates because of the diseases or road accidents due to low visibility

  It was found that there was a positive correlation for respiratory diseases and mortality rate with a statistically significant p value. This tells us that we can conclude based on the available data, there has been an increased risk to life for people as the smoke fire has increased over the years. For road accidents, the correlation between road accidents and smoke fires was negative. However, this was not statistically significant, which makes these results unreliable.

- Which respiratory diseases are shown to have a significant increase in mortality rates

  Chronic Respiratory diseases and Chronic obstructive pulmonary disease were shown to have the biggest increase in mortality rates with the increase in smoke fires in Derby, Kansas.

Demographic Disparity

- Identifying age groups or ethnicities most vulnerable or impacted by health conditions.

  It was seen the rise in mortality rates was noticeably higher in females for Chronic Respiratory diseases and Chronic obstructive pulmonary disease as compared to females. Also, the mortality rates for cardiovascular diseases were shown to have a greater impact on AIAN (American Indian and Alaska Native) ethnicity. This was validated through the ANOVA test which showed that there was a significant difference between average mortality rates for different ethnic groups. Also, it was observed that mortality rates for road accidents were significantly higher for seniors and young adults . There was also a visible local trend where the mortality rates due to road accidents increased in the years when there was an increase in smoke. However, the correlation on an overall level came out to be negative, which was not statistically significant.

- Observe change in life expectancy across different demographics.

  The AIAN (American Indian and Alaska Native) has seen a noticeable decline in the life expectancy at birth as compared to any other ethnic group. This was also noticeable through the increasing mortality rates for AIAN across different diseases.

The importance of Human Centered Approach to Data Science can be seen in this analysis where it analyzed the impact of smoke fires by understanding the perspective of residents in Derby who are directly affected by the fires, communicating the findings to them in way they understand, and including everyone in the analysis to ensure no misrepresentation. The findings can then be used to make personalized health plans, improve driving guidelines, and empower every community. This approach ensures that the analysis is fair and ethical, incorporates feedback from the community, and allocates resources in the best way to make life better for those impacted by smoke fires.

References

1. Rice MB, Henderson SB, Lambert AA, Cromar KR, Hall JA, Cascio WE, Smith PG, Marsh BJ, Coefield S, Balmes JR, Kamal A, Gilmour MI, Carlsten C, Navarro KM, Collman GW, Rappold A, Miller MD, Stone SL, Costa DL. Respiratory Impacts of Wildland Fire Smoke: Future Challenges and Policy Opportunities. An Official American Thoracic Society Workshop Report. Ann Am Thorac Soc. 2021 Jun;18(6):921-930. doi: 10.1513/AnnalsATS.202102-148ST. PMID: 33938390; PMCID: PMC8456726.
2. Reid CE, Jerrett M, Tager IB, Petersen ML, Mann JK, Balmes JR. Differential respiratory health effects from the 2008 northern California wildfires: A spatiotemporal approach. Environ Res. 2016 Oct;150:227-235. doi: 10.1016/j.envres.2016.06.012. Epub 2016 Jun 15. PMID: 27318255.
3. Hoshiko S, Buckman JR, Jones CG, Yeomans KR, Mello A, Thilakaratne R, Sergienko E, Allen K, Bello L, Rappold AG. Responses to Wildfire and Prescribed Fire Smoke: A Survey of a Medically Vulnerable Adult Population in the Wildland-Urban Interface, Mariposa County, California. Int J Environ Res Public Health. 2023 Jan 10;20(2):1210. doi: 10.3390/ijerph20021210. PMID: 36673971; PMCID: PMC9858942.
4. Intini, Paolo & Wahlqvist, Jonathan & Wetterberg, Niklas & Ronchi, Enrico. (2022). Modelling the impact of wildfire smoke on driving speed. International Journal of Disaster Risk Reduction. 80. 103211. 10.1016/j.ijdrr.2022.103211.

Data Sources

1. https://ghdx.healthdata.org/record/ihme-data/united-states-chronic-respiratory-disease-mortality-rates-county-1980-2014
2. https://ghdx.healthdata.org/record/ihme-data/united-states-causes-death-life-expectancy-by-county-race-ethnicity-2000-2019
3. https://www.sciencebase.gov/catalog/item/61aa537dd34eb622f699df81
4. US City assigned for individual analysis - Google spreadsheet

The above data sources included most of the information that was needed to address the research questions and validate the statistical significance of our results. The data is available on the GHDx website which is the world's most comprehensive catalog of surveys, censuses, vital statistics, and other health-related data. The data is open source and widely accessible to the users and contains all the information available for any healthcare analysis. THE USGS data contains the Combined wildland fire datasets for the United States and certain territories in the json format.

Data Profiling

Data Source 1

The data contains Mortality rates for different counties in Kansas with their FIPS between the time period 1980 - 2014. Mortality rates for different counties in Kansas where mortality rate is deaths per 100,000 population.

FIPS - The Federal Information Processing Standards (FIPS) code, a unique identifier for states and counties in the United States

Data is available for the following respiratory conditions -

- Chronic respiratory diseases
- Chronic obstructive pulmonary

- Pneumoconiosis
- Silicosis
- Asbestosis
- Coal workers pneumoconiosis
- Other pneumoconiosis
- Asthma
- Interstitial lung disease
- Other chronic respiratory

This data source helps address the first question which focuses on finding a correlation between smoke fires and overall mortality rates due to respiratory conditions. The analysis would further review if there are certain respiratory diseases which have seen a major increase as compared to others and how their correlation is with increase in smoke fires.

Data Source 2

This data contains a detailed view of the mortality rates due to respiratory conditions, cardiovascular health and road accidents with columns such as Location, FIPS, Race, Sex, Age group, Cause of death between the time period 2000-2019. The estimates are given in the following three columns :

Val - Estimate for that specific county	Lower - 2.5% percentile estimate	Upper - 97.5% percentile estimate

The data also contains life expectancy by different demographic to validate the change in mortality rates

This data source will help find patterns across different demographics to identify highly prone risk groups who are being affected the most. The analysis would initially be done at a univariate level for each demographic data such as race, sex and age.

Data Source 3

The USGS data contains the combined wild fire datasets for the United States and certain territories. The data has been collected by the US Geological Survey and is well documented. The data is available in a few formats and the analysis has been conducted using the GeoJSON format.